2019-04-30

Data Enhancement and Analysis of Federal Acquisition Databases

Wu, Ningning; Tudoreanu, M. Eduard; Wang, Richard; Jiang, Wenxue

Monterey, California. Naval Postgraduate School

SYM-AM-19-057

PROCEEDINGS
OF THE
SIXTEENTH ANNUAL ACQUISITION RESEARCH SYMPOSIUM

WEDNESDAY SESSIONS
VOLUME I

Acquisition Research:
Creating Synergy for Informed Change

May 8–9, 2019

Published: April 30, 2019

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Data Enhancement and Analysis of Federal Acquisition Databases

Ningning Wu—is Professor of Information Science at the University of Arkansas at Little Rock. She received a BS and an MS in Electrical Engineering from the University of Science and Technology of China and PhD in Information Technology from George Mason University. Wu's research interests are data mining, network and information security, and information quality. She holds certificates of the IAIDQ Information Quality Certified Professional (IQCP) and the SANS GIAC Security Essentials Certified Professional. [nxwu@ualr.edu]

M. Eduard Tudoreanu—is Professor of Information Science at University of Arkansas Little Rock. Tudoreanu has expertise in human-computer interaction, information quality, advanced visualization of complex data, and virtual reality. He worked on visual data analysis, and has extensive experience in software development and user interface design. [metudoreanu@ualr.edu]

Richard Wang—is Director of the MIT Chief Data Officer and Information Quality Program. He is also the Executive Director of the Institute for Chief Data Officers (iCDO) and Professor at the University of Arkansas at Little Rock. From 2009 to 2011, Wang served as the Deputy Chief Data Officer and Chief Data Quality Officer of the U.S. Army. He received his PhD in information Technology from the MIT Sloan School of Management in 1985. [rwang@mit.edu]

Wenxue Jiang—is a graduate student in the Information Quality program at University of Arkansas at Little Rock. He has been working on his master project with a focus on quality assessment and integration of acquisition databases. He is expected to graduate with a Master of Science in Information Quality in December 2019. [wxjang@ualr.edu]

Abstract

The Federal Funding Accountability and Transparency Act of 2006 (FFATA) required federal contract, grant, loan, and other financial assistance awards of more than $25,000 be displayed on a publicly accessible and searchable website to give the American public access to information on what the federal government spends every year and how it spends the money. Federal acquisition databases, such as those maintained by usaspending.gov and fpds.gov, serve this purpose. These databases contain contract information for all U.S. departments for the last 20 years. However, little has been done to dig into the data and extract the information that may provide valuable insights on potential ways to improve the efficiency of acquisition management. This paper takes a data science approach to assessing and enhancing the quality of the databases and to discovering patterns that can be potentially useful for acquisition research and practice.

Introduction

Defense acquisition consists of different data silos. These data silos have both technical and cultural origins. The capabilities to draw upon data across information systems hold huge potential for improving defense acquisition and procurement. Acquisition planning and management involves many decision-making and action-taking processes that cover a complex environment including actual acquisition, contracting, fiscal, legal, personnel, and regulatory requirements. A sound decision-making process has to rely on data—high quality data. Often the available data is dirty, outdated, incomplete, or insufficient for the expert to make a decision. On the other hand, there are enormous amounts of data on the web that can be utilized to crystalize the needed information.

The paper will investigate how to leverage the information in public data sources to complement the internal data in order to support effective acquisition planning and management. The research is based on publicly accessible government acquisition

databases at usaspending.gov and fpds.gov. Both databases host federal spending data from the last two decades and contain millions of records with detailed information about each contract. These rich repositories of data provide a great opportunity for us to learn from the past practices, and, hopefully, to gain some insights that can help us design better strategies for managing future projects.

A preliminary study showed that the acquisition data suffer from the quality problems as do all other real-world data. To achieve high quality data analytics, we have to improve the quality of data. Our previous research demonstrated the feasibility of using online information from reputable sources to fill the missing values and correct erroneous or inconsistent data of acquisition databases. The research in this paper takes that a step further. It aims to enhance the acquisition data with online information so as to discover patterns that otherwise would not be able to be found.

Trust is a key issue for using online data. In fact, the web has not only changed our ways of sharing and seeking information, it has also altered traditional notions of trust due to the fact that the information can be published anywhere by anyone for any purpose, and there is no authority to certify the correctness of the information. It is often up to the information consumers to make their own judgement about the credibility and accuracy of information they encountered online. Unfortunately, in the world nowadays, people are flooded with fake news and internet scams. Thus it becomes even harder for an information seeker to discriminate between true and false information. To make the situation even worse, even when data are deemed trustworthy, assessing the data quality in this big data era still brings many challenges. First, the diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration. Second, data change very fast and the timeliness of data is very short, which necessitates higher requirements for processing technology (Cai & Zhu, 2015).

This paper explores only the usage of information from credible and reputable sources to enhance the data analytics ability. However, investigating appropriate methods to assess web data quality, to identify and acquire credible and accurate information will be one of our future research topics.

Research Methodology

The research work follows the Data Enhancement and Analytics System framework shown as Figure 1 (Wu, Tudoreanu, & Wang, 2018).

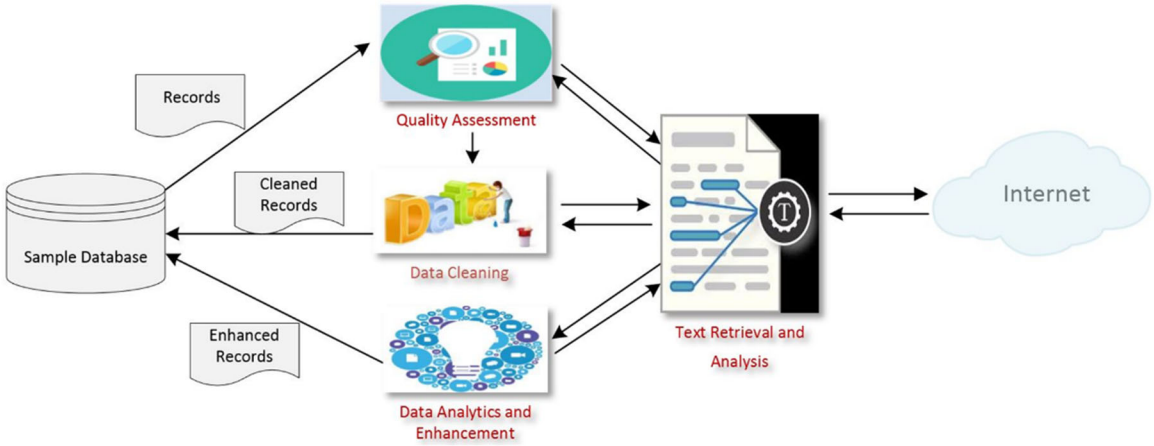

Figure 1. Framework of Data Enhancement and Analytics System

Our research methodology contains the following steps:

- Compare the data between fpds.gov and usaspending.gov in terms of their structures, contents, and quality.
- Apply data analytics techniques to discover patterns about past acquisition projects. These patterns might help us to identify the room for improvement in future projects.

Comparison of FPDS and USASPENDING Data

Both usaspending.gov and fpds.gov sites are publicly accessible and have the contract information of all U.S. departments since 2000; however, the data in two sites are organized in different structures with a different number of attributes. The data in usaspending.gov are categorized under prime award and sub-award. The types of spending include contracts, grants, loans, and other financial assistance. For each spending type, the data is organized into two structures: prime award and sub-award. For example, information on contracts is organized into two tables: one for prime contracts and the other for sub contracts. Data in fpds.gov is organized using a unified structure. We downloaded the spending data of the Department of Defense and stored them on a MYSQL database server.

Table 1 shows the structure of tables from each website, where the fpds row is from fpds.gov, and the other rows are from usaspending.gov. Here, RecCnt and ColCnt represent the number of records and number of columns in a table respectively; CompleteCols and SingleValCols represent the number of columns with no missing values and number of columns with only a single value across all records; and EmptyCols and IncompleteCols represent the number of empty columns and the number of columns with missing values respectively.

Table 1. Profiling of FPDS and usaspending Tables

Table Name	ColCnt	CompleteCols/ SingleValCols	EmtpyCols	IncompleteCols
PrimeContracts	221	50/1	0	162
SubContracts	101	41/0	3	57
PrimGrants	67	32/5	2	33
SubAGrants	101	29/4	25	47
fpds	210	74/3	1	136

A close study of these tables reveals that the fpds table is similar to the PrimeContracts table from usaspending.gov in terms of their contents. Thus, the remaining part of this section compares only these two tables in terms of their schema, data coverage, and quality.

To facilitate the data comparison, attributes are classified into two categories: identity attributes and non-identity attributes. Identity attributes provide identity information for a contractor, contract, funding agency, etc. Examples of identity attributes include project identifier, contractor identifier (such as a DUNS number), business name, address information, phone, fax, etc. Non-identity attributes do not provide any identity information.

Attribute Naming Convention

PrimeContracts uses key description abbreviation to construct attribute names. Fpds groups attributes into categories. It then uses a key descriptor plus a category prefix to

name an attribute. Compared to the PrimeContracts table, fpds attributes have longer but easy-to-understand names. The fpds attribute categories and the number attributes for each category are shown as in Figure 2.

- **awardID:** 7 cols
- **competition:** 20 cols
- **contractData:** 25 cols
- **contractMarketingData:** 1 cols
- **dollarValues:** 3 cols
- **legislativeMandates**: 10 cols
- **placeOfPerformance**: 5 cols
- **performancePrograms:** 1 cols
- **productOrServiceinformation:** 11 cols
- **purchaserInformation:** 5 cols
- **relevantContracDates:** 4 cols
- **transactionInformation:** 8 cols
- **vendor:** 110 cols

Figure 2. FPDS Attribute Categories

Schema Mapping

Schema mapping between the two tables are performed manually based on the data dictionary provided by each database. There are 180 common fields in the two tables even though these fields are named differently in each table. The remaining 30 attributes in fpds and 41 attributes in PrimeContracts are found only in their own table. Due to space limitations, Table 2 only shows the partial mapping results.

Table 2. Schema Mapping Between fpds and PrimeContracts Tables

(a) Mapping of Common Attributes

	Mapping Attributes	
	Attributes in fpds	**Matched Attributes in PrimeContracts**
1	awardID_awardContractID_PIID	piid
2	awardID_awardContractID_agencyID	agencyid
3	awardID_awardContractID_modNumber	modnumber
4	awardID_awardContractID_transactionNumber	transactionnumber
5	awardID_referencedIDVID_PIID	idvpiid
6	awardID_referencedIDVID_agencyID	idvagencyid
7	awardID_referencedIDVID_modNumber	idvmodificationnumber
8	competition_A76Action	a76action
9	competition_commercialItemAcquisitionProcedures	commercialitemacquisitionprocedures
10	competition_commercialItemTestProgram	commercialitemtestprogram
11	competition_competitiveProcedures	competitiveprocedures
12	competition_evaluatedPreference	evaluatedpreference
13	competition_extentCompeted	extentcompeted
14	competition_fedBizOpps	fedbizopps
15	competition_idvNumberOfOffersReceived	numberofoffersreceived
	.....	.....
	.....	.....
165	vendor_vendorSiteDeta__ndorSocioEconomicIndicators_isIndianTribe	isindiantribe
166	vendor_vendorSiteDeta__allyDisadvantagedWomenOwnedSmallBusiness2	isecondisadvwomenownedsmallbusiness
167	vendor_vendorSiteDeta__ors_isJointVentureWomenOwnedSmallBusiness	isjointventurewomenownedsmallbusiness
168	vendor_vendorSiteDeta__s_isNativeHawaiianOwnedOrganizationOrFirm	isnativehawaiianownedorganizationorfirm
169	vendor_vendorSiteDeta__erviceRelatedDisabledVeteranOwnedBusiness	srdvobflag
170	vendor_vendorSiteDeta__cioEconomicIndicators_isTriballyOwnedFirm	istriballyownedfirm
171	vendor_vendorSiteDeta__dorSocioEconomicIndicators_isVeteranOwned	veteranownedflag
172	vendor_vendorSiteDeta__endorSocioEconomicIndicators_isWomenOwned	womenownedflag
173	vendor_vendorSiteDeta__nomicIndicators_isWomenOwnedSmallBusiness	iswomenownedsmallbusiness
174	vendor_vendorSiteDeta__Owned_isAsianPacificAmericanOwnedBusiness	apaobflag
175	vendor_vendorSiteDeta__inorityOwned_isBlackAmericanOwnedBusiness	baobflag
176	vendor_vendorSiteDeta__rityOwned_isHispanicAmericanOwnedBusiness	haobflag
177	vendor_vendorSiteDeta__cIndicators_minorityOwned_isMinorityOwned	minorityownedbusinessflag
178	vendor_vendorSiteDeta__norityOwned_isNativeAmericanOwnedBusiness	naobflag
179	vendor_vendorSiteDeta__cators_minorityOwned_isOtherMinorityOwned	isotherminorityowned
180	vendor_vendorSiteDeta___isSubContinentAsianAmericanOwnedBusiness	saaobflag

(b) Unique Attributes of Each Table

	Unique Attributes	
	Unique Attributes in fpds	Unique Attributes in PrimeContracts
1	competition_idvTypeOfSetAside	congressionaldistrict
2	competition_numberOfOffersReceived	divisionnumberorofficecode
3	competition_numberOfOffersSource	emergingsmallbusinessflag
4	competition_typeOfSetAsideSource	fiscal_year
5	contractData_inherentlyGovernmentalFunction	hubzoneflag
6	contractData_listOfTreasuryAccounts_treasuryAccount_initiative	isarchitectureandengineering
7	contractData_listOfTr__yAccounts_treasuryAccount_obligatedAmount	isconstructionfirm
8	contractData_listOfTr__nt_treasuryAccountSymbol_agencyIdentifier	isotherbusinessororganization
9	contractData_listOfTr__unt_treasuryAccountSymbol_mainAccountCode	isserviceprovider
10	contractData_listOfTr__ount_treasuryAccountSymbol_subAccountCode	lastdatetoorder
11	contractData_undefinitizedAction	lettercontract
12	contractMarketingData_feePaidForUseOfService	locationcode
13	legislativeMandates_constructionWageRateRequirements	maj_agency_cat
14	legislativeMandates_laborStandards	maj_fund_agency_cat
15	legislativeMandates_l__lReportingValues_additionalReportingValue	mod_agency
16	legislativeMandates_materialsSuppliesArticlesEquipment	mod_parent
17	transactionInformation_closedBy	multipleorsingleawardidc
18	transactionInformation_closedDate	parentdunsnumber
19	transactionInformation_closedStatus	pop_cd
20	transactionInformation_createdBy	prime_awardee_executive1
21	transactionInformation_createdDate	prime_awardee_executive1_compensation
22	**transactionInformation**_lastModifiedBy	prime_awardee_executive2
23	vendor_vendorHeader_vendorAlternateName	prime_awardee_executive2_compensation
24	vendor_vendorSiteDeta__rtifications_isSBACertified8AJointVenture	prime_awardee_executive3
25	vendor_vendorSiteDeta__endorCertifications_isSBACertifiedHUBZone	prime_awardee_executive3_compensation
26	vendor_vendorSiteDeta__ations_isSelfCertifiedHUBZoneJointVenture	prime_awardee_executive4
27	vendor_vendorSiteDetails_vendorDUNSInformation_cageCode	prime_awardee_executive4_compensation
28	vendor_vendorSiteDeta__rganizationFactors_**countryOfIncorporation**	prime_awardee_executive5
29	vendor_vendorSiteDeta__rOrganizationFactors_stateOfIncorporation	prime_awardee_executive5_compensation
30	vendor_vendorSiteDeta__cioEconomicIndicators_isVerySmallBusiness	programacronym
31		progsourceaccount
32		progsourceagency
33		progsourcesubacct
34		psc_cat
35		rec_flag
36		statecode
37		streetaddress3
38		typeofidc
39		unique_transaction_id
40		vendorenabled
41		vendorlocationdisableflag

Quality Assessment

Due to the space limitation, only the quality assessment of key identity attributes is presented here. Quality assessment is performed on the dimensions of column completeness, and field length consistency of attributes that have fixed-length values. Table 3 shows that the fpds table has a higher column completeness measure than the PrimeContracts table. Figures 3 and 4 show the field length distribution of the PIID (prime project ID) and prime contractor DUNS numbers respectively. Since the PIID is a system wide identifier for each prime project, it is assumed to have a fixed length. But there are some exceptions in both the fpds and PrimeContract tables. Similarly, the DUNs number is a 9-digit value. Any DUNS numbers other than 9-digit are considered incorrect.

Table 3. Column Completeness

Table Name	ColCnt	IncompleteCols	%CompleteCols
PrimeContracts	212	162	23.6%
fpds	210	136	35.2%

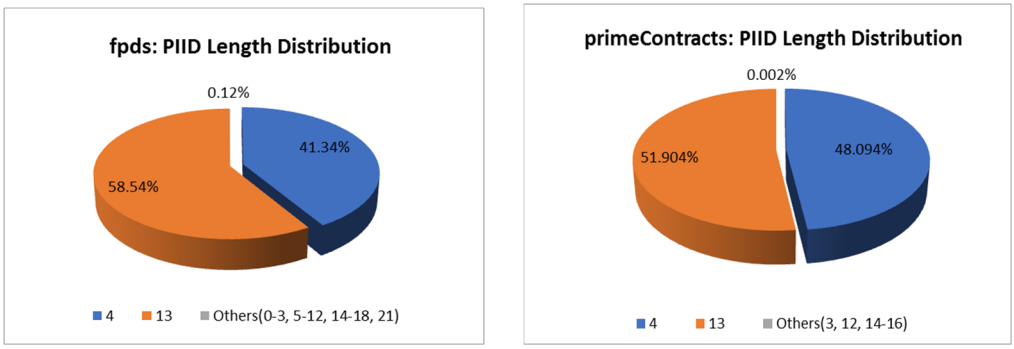

Figure 2. PIID Length Distribution

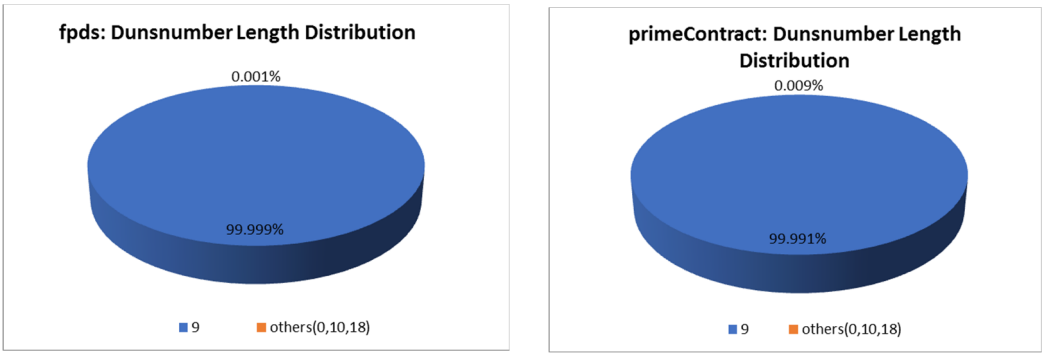

Figure 3. DUNs Number Length Distribution

Record Mapping

Record mapping matches records of the two tables if they represent the same entity. In fpds and PrimeContracts, each contract is considered as an entity. Since both tables contain the contract information from the Department of Defense, record mapping provides a way to measure the data consistency between them. Record mapping is a typical entity resolution process. It requires comparing fields of records to determine whether they belong to the same entity or not. If records have common key identifier attributes, mapping them is rather straightforward; otherwise, the non-identifier attributes have to be used to determine how similar the records are. Unfortunately, the fpds and PrimeContracts tables don't have a common record identifier, thus record mapping must rely on the common attributes of two tables.

Considering the number of attributes and records in the fpds and PrimeContracts tables, record mapping is a very complicated and time-consuming process. Thus, the first phase of mapping is performed on sample data instead, and it considers only the following identity attributes when matching records: PIID, dunsnumber, vnedorlocationzipcode, vendorlocationstate, vendorlocationcity, vendor_countrycode, vendor_phoneno, and vendorlocation_streetaddress. Here, PIID denotes the primary project ID that is unique to each project. Dunsnumer denotes the 9-digit DUNS number of the primary contractor of a project. vnedorlocationzipcode, vendorlocationstate, vendorlocationcity, vendor_countrycode, vendor_phoneno, and vendorlocation_streetaddress represent address and telephone information of a primary contractor. Two records are considered to represent the same entity if their values on each of the above attributes match.

The following steps are performed to prepare the sample datasets:

- A random sample of 5000 PIIDs that exist in both tables is drawn.

- The corresponding records of these PIIDs are retrieved from the fpds and PrimeContract tables respectively and they are stored into separate datasets, denoted as datasets D_f, and D_u.
- As data quality issues will adversely affect the record matching result, data standardization and transformation are performed. Duplicate records and records with missing values are removed.
- The equijoin is applied on two datasets, and the resulting dataset is denoted as D_{join}.

Figure 4 compares the number of distinct values of each identity attribute among three datasets D_f, D_u, and D_{join}. It shows that D_u consistently has more distinct values for each attribute than D_f. The number of distinct values for each attribute in table D_{join} indicates the number of attribute common values between D_f and D_u.

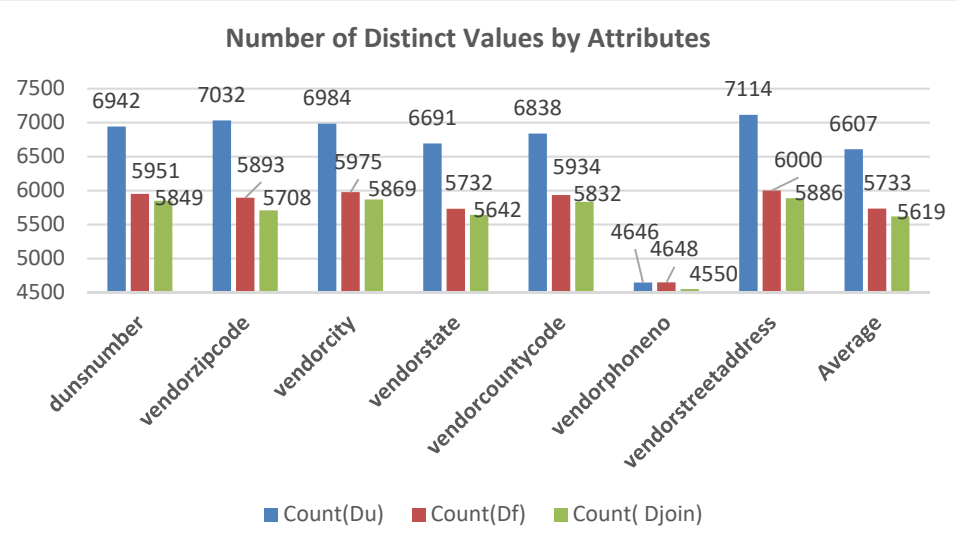

Figure 4. Number of Distinct Values by Attributes

Figure 5 shows the relative consistency measure of each attribute of one table in terms of the other table. For example, 98.3% of dunsnumbers in D_f are also found in D_u, while only 84.3% of dunsnumber in D_u are found in D_f; 96.7% of vendorzipcodes in D_f are also found in D_u, but 81.2% of vendorzipcodes in D_u are found in D_f. The reason behind these discrepancies is that, given a prime award ID, there are more distinct records in D_u than in D_f. Possible root causes may include the following: fpds.gov and usaspending.gov collected the data at different granularity levels, the fpds database may miss some records, or the usaspending database may have to keep multiple records for the same prime award as these records have inconsistent values and it is not clear which values are right and which are not.

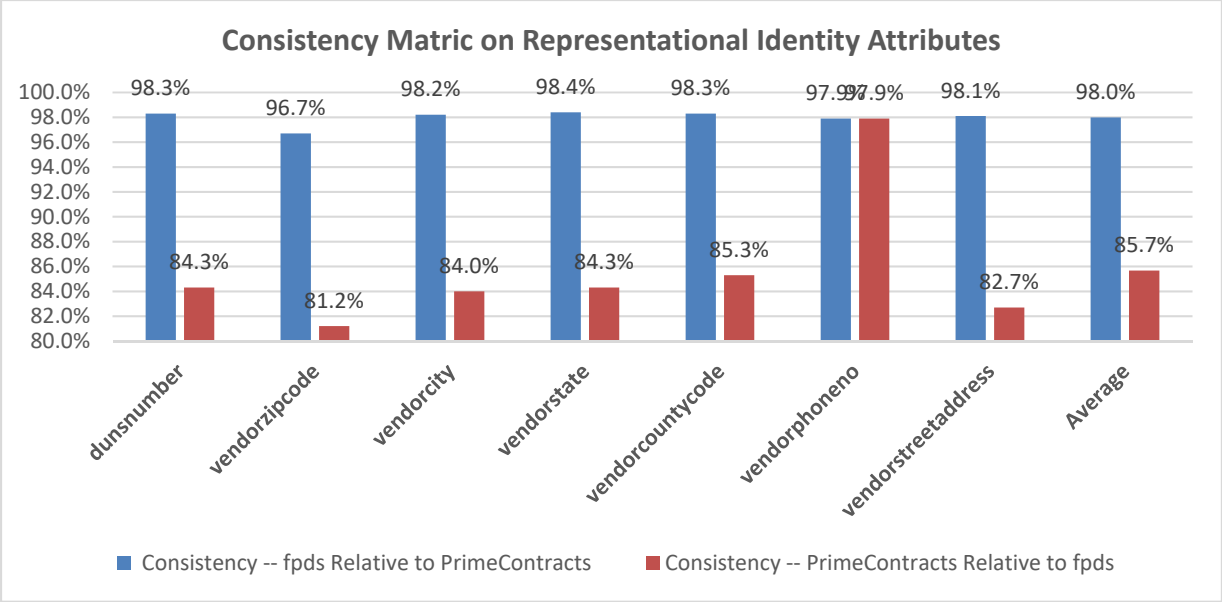

Figure 5. Relative Consistency Measure of Each Attribute

Data Analytics

The goal of data analytics is to discover hidden and interesting patterns that can be potentially useful in planning future acquisition projects. Since we are not the domain expert on acquisition data and policies, we decide to take data science approach and start the data analytics with a hypothesis.

Hypothesis 1: Critical contractors are those that provide unique products and services. They could be the weakest link in a supply chain, because if they failed, it would be hard to find alternatives to fill their places.

North American Industry Classification System (NAICS) is the standard used by federal statistical agencies in classifying business establishments for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy. NAICS code describes the business specialization of a company.

There are 379 distinct NAICS codes among all contractors. Seventy-eight NAICS codes have only one contractor associated with it. This means in the current pool of DoD contractors, these 78 contractors are critical contractors as no other DoD contractors are doing the same business. It is possible that there are companies that, outside the DoD contractor pool, are associated with these NAICS codes. On average, each of those critical contractors is involved in 37 different projects. The top 10 critical contractors with the most number of projects is listed in Table 4.

Table 4. The Top 10 Critical Contractors With the Most Number of Projects

Rank	No. of Distinct Projects
1	399
2	382
3	343
4	245
5	237
6	138
7	117
8	91
9	69
10	61

For those highly demanded contractors, most of them are big and well-established companies, but a couple of them are small companies that appear to provide very unique products and services. These companies could be a potential weak point in a project/supply chain and may critically affected the overall outcome of a project if they fail.

Hypothesis 2: A primary project usually has hundreds of contractors working on it. These contractors spread out in different geographical locations. Some may be located in an area with a high risk of natural disasters such as earthquakes, flooding, hurricanes, tornados, and so forth. Some natural disasters, like tornados and earthquakes, are hard to predict. Thus, it would be always beneficial to consider those risk factors when planning a project. Possible strategies include using contractors located in low-risk areas, or intentionally selecting contractors that are spread out in different geographical locations, or having backup plans in place to handle any emergencies.

We have obtained the natural disaster data for each U.S. county between the years 1950 and 2018 from the National Centers for Environmental Information (Formerly the National Climatic Data Center [NCDC]). The data cover all types of natural disasters, including floods, tornados, hurricanes, blizzards, high winds, flash floods, hail, dust storms, and so forth.

This project focuses on disasters that could cause severe damages and significantly affect the normal life and business operations of local communities such as tornados, hurricanes, floods, and blizzards. Since the world weather has changed quite fast in recent decades, we decided to use the NCDC data of last 20 years to identify whether an area is prone to a natural disaster based on the following criteria. The high-risk flooding areas are identified as those that have at least 10 episodes of floods in the last 20 years; the high-risk hurricane areas are those that have at least one hurricane in last 20 years; the high-risk wildfire areas are those that have at least one wildfire that lasted more than one day in last 20 years; and the high-risk tornado areas are those that have at least one category 3 or above tornado in the last 20 years. Table 5 shows the number of subcontractor zip codes belong to each disaster type.

Table 5. Number of Subcontractor Zip Codes Vulnerable to Each Disaster Type

Disaster Type	Flood	Hurricane	Tornado	Wildfire
# zipcodes	5959	780	1182	1831

Our analysis found that there are 6,786 natural disaster–prone zip codes of the principal places where the work is performed for a subcontract. Some of these zip codes are vulnerable to more than one disaster type. The natural disaster–prone areas are further categorized into four classes based on the number of distinct disaster types that has been observed in that area during the last 20 years.

Table 6 shows the distribution of subcontract principal place zip by the number of disaster types along with the distribution of subcontractors located in those zip codes. The column %zip_population indicates the percentage of zip codes (of a category) with regard to the total number of subcontract zip codes, and %DUNS_population indicates the percentage of DUNs in each category of zips with regard to total number of subcontractor DUNS number.

Table 6. Distribution of Subcontractor Principal Zip and DUNS

#DisasterTypes	#zipcodes	%zip_population	#duns	% DUNS_population
1	2165	7.8%	13373	42.3%
2	3548	12.9%	10965	34.6%
3	1004	3.6%	2733	8.6%
4	69	0.25%	141	0.44%
Total:	6786	23.7%	27072	86.0%

Subcontractors that are located in an area vulnerable to all four disaster types are considered to have a high risk. Table 7 shows the top 10 projects with the highest number of high-risk contractors.

Table 7. Top 10 Projects With the Highest Number of High-Risk Contractors

Rank	No. of High-Risk Contractors
1	59
2	49
3	43
4	37
5	36
6	31
7	27
8	24
9	23
10	19

It would be interesting to know the percentage of high-risk contractors in past projects. There are total 588 projects have at least one high-risk subcontractors. Figure 6 shows the distribution of projects by their percentage of contractors that are vulnerable to all four types of natural disaster. A close study reveals that the majority of 129 projects in the last bin with more than 90% of subcontractors in high-risk areas have only one subcontractor. More than half of 588 projects have less than 10% of subcontractors in high-risk areas. Ideally, a project should have as few as possible high-risk subcontractors.

We believe the information on high-risk areas of natural disasters is beneficial because it helps project managers calculate the risk of a project and develop strategies to mitigate the risk to the minimum.

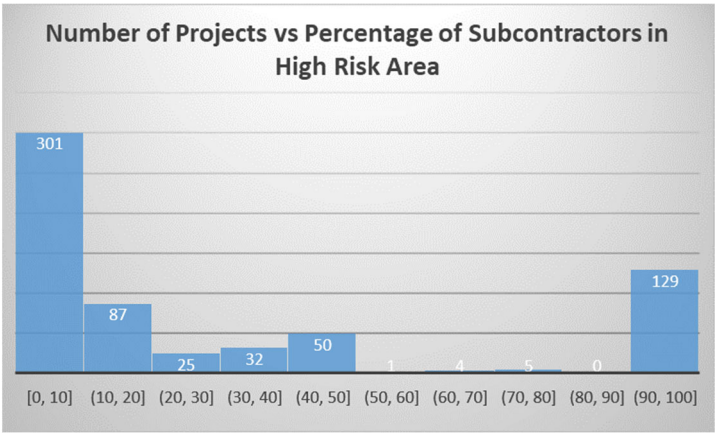

Figure 6. Distribution of Projects by Percentage of High-Risk Subcontractors

Related Work

This section summarizes some related work in the fields of federal acquisition data analysis.

Tudoreanu et al. (2018) investigated employment data in an attempt to correlate changes in employment with negative modifications to contracts. Such correlations can be explored to infer hidden and undisclosed contractors. Hidden contractors may pose the risk of becoming a weak, stress point of a project and would affect the overall outcome of the project.

Wu et al. (2018) proposed a framework based on data science approach that aims to utilize the online information to assess and improve acquisition database quality as well as to find the hidden patterns to further acquisition research. The main component of the framework is a web-search and text mining module, whose main function is to search the internet and identify the most credible and accurate information online.

Apte, Rendon, and Dixon (2015) explored the use of Big Data analytic techniques to explore and analyze large dataset that are used to capture information about DoD services acquisitions. The paper described how big data analytics could potentially be used in acquisition research. As the proof of concept, the paper tested the application of Big Data Analytic techniques by applying them to a dataset of Contractor Performance Assessment Report System (CPARS) ratings of 715 acquired services. It also created predictive models to explore the causes of failed services contracts. Since the dataset used in the research was rather small and far from the scope of big data, the techniques explored by the paper mainly focus on traditional data mining techniques without taking into account big data properties.

Black, Henley, and Clute (2014) studied the quality of narratives in CPARS and their value to the acquisition process. The research used statistical analysis to examine 715 Army service contractor performance reports in CPARS in order to understand three major questions: (1) To what degree are government contracting professionals submitting to CPARS contractor performance narratives in accordance with the guidelines provided in the CPARS user's manual? (2) What is the added value of the contractor performance narratives beyond the value of the objective scores for performance? (3) What is the statistical relationship between the sentiment contained in the narratives and the objective scores for contractor evaluations?

Conclusion and Future Work

This research presented a data science approach to compare and analyze publicly accessible acquisition databases. The research explored the usage of online information to enhance the internal data in order to discover the hidden patterns in the data. The research has collected natural disaster information from the National Centers for Environmental Information. This information can be helpful in identifying high-risk locations and contractors located in those locations.

Future work will focus on the following two directions. First, explore more data analytics techniques to discover patterns that are potentially useful to the acquisition research community. Second, research effective text mining techniques for assessing web data quality and retrieving credible information from online sources.

References

Apte, U., Rendon, R., & Dixon, M. (2016). Big data analysis of contractor performance information for service acquisition in DoD: A proof of concept. In *Proceedings of the 13th Annual Acquisition Research Symposium*. Monterey, CA: Naval Postgraduate School.

Augustine, N. R. (1997). *Augustine's laws*. AIAA.

Black, S., Henley, J., & Clute, M. (2014). *Determining the value of Contractor Performance Assessment Reporting System (CPARS) narratives for the acquisition process* (NPS-CM-14-022). Monterey, CA: Naval Postgraduate School.

Brown, B. (2010). *Introduction to defense acquisitions management.* Fort Belvoir: VA: Defense Acquisition University. Retrieved from www.dau.mil/publications/publicationsDocs/Intro%20to%20Def%20Acq%20Mgmt%2010%20ed.pdf

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal, 14*, 2. doi: http://doi.org/10.5334/dsj-2015-002

Cheskin, S. (1999). *Ecommerce trust: Building trust in digital environments*. Archetype/Sapient.

Cilli, M., Parnell, G. S., Cloutier, R., & Zigh, T. (2015). A systems engineering perspective on the revised defense acquisition system. *Systems Engineering, 18*(6), 584–603. doi:10.1002/sys.21329.

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies*, *58*(6), 737–758.

DAU. (n.d.). DAU Center for Defense Acquisition Research agenda 2016–2017. Retrieved from http://dau.dodlive.mil/files/2016/01/ARJ-76_ONLINE-FULL.pdf

DoD. (2007, November). *Operation of the Defense Acquisition System* (DoDI 5000.01). Washington, DC: Author.

DoD. (2015). *Operation of the Defense Acquisition System* (DoDI 5000.02). Washington, DC: Author.

Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., … Treinen, M. (2001). What makes web sites credible?: A report on a large quantitative study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 61–68). New York, NY: ACM Press.

Gallup et al. (2015, May). Lexical Link Analysis (LLA) application: Improving web service to defense acquisition visibility environment. *Distributed Information Systems Experimentation*.

Gaither, C. C. (2014). Incorporating market based decision making processes in defense acquisitions. *International Journal of Defense Acquisition Management, 6*, 38–50.

Golbeck, J. (2008). Trust on the world wide web: A survey. *Foundations and Trends® in Web Science, 1*(2), 131–197.

Hagan, G. (1998). *Glossary: Defense acquisition acronyms and terms*. Fort Belvoir, VA: DoD, Defense Systems Management College, Acquisition Policy Department.

Krzysko, M. (2012, February). The need for acquisition visibility. *Journal of Software Technology*, 4–9.

Krzysko, M. (2016). *Acquisition decision making through information and data management.* Retrieved from www.digitalgovernment.com/media/Downloads/asset_upload_file917_5737.pdf

McKernan, M., Moore, N. Y., Connor, K., Chenoweth, M. E., Drezner, J. A., Dryden, J., … Szafran, A. (2016). *Issues with access to acquisition and information in the Department of Defense.* Santa Monica, CA: Rand Corporation.

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics, 59*, 210–220.

Miller, A., & Ray, J. (2015, January). Moving from standard practices to best practices in defense acquisition. *Defense ARJ, 22*(1), 64–83.

Pennock, M. J. (2008). Defense acquisition: A tragedy of the commons. Retrieved from *ProQuest*.

Tudoreanu, M. E., Franklin, K., Wu, N., & Wang, R. (2018). Searching hidden links: Inferring undisclosed subcontractors from public contract records and employment data. In *Proceedings of the 15th Annual Acquisition Research Symposium*. Monterey, CA: Naval Postgraduate School.

Wu, N., Tudoreanu, M. E., & Wang, R. (2018). Leveraging public data for quality improvement and pattern discovery of federal acquisition data. In *Proceedings of the 15th Annual Acquisition Research Symposium*. Monterey, CA: Naval Postgraduate School.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CA 93943

www.acquisitionresearch.net